Assistance dog candidate selection based on temperament: the reactivity–stability dimension in the first year of life

Zwaan Pijnenburg

Assistance dog candidate selection based on temperament: the reactivity–stability dimension in the first year of life

by

Zwaan Pijnenburg
Student ID 1017615

MSc Thesis Behavioural Ecology
Course Code BHE80336

Under the Supervision of
prof. dr. M (Marc) Naguib
J R (Jori) Noordenbos MSc

Submitted to
Behavioural Ecology
Wageningen University

In Partial Fulfilment of the Requirements for the Degree of
MSc Biology

16 November 2025

Abstract

Assistance dogs help people with disabilities navigate daily life. Dogs are selected for assistance dog training based on temperament. Ideally, behavioural assessments are conducted at the youngest age possible to save resources and promote welfare by releasing dogs with undesirable temperaments. One temperament dimension of interest is reactivity–stability, defined as the response to novel, unexpected stimuli. This thesis investigated the predictive validity of reactivity–stability for qualification as an assistance dog trainee and assignment to specific roles, based on behavioural observations from three startle tests in the behavioural assessment currently in use at Hulphond Nederland, a Dutch assistance dog programme. Dogs were exposed to stimuli that elicit reactivity in an acoustic startle, novel object startle, and moving object startle test. Thirty-four subjects were tested at 13 months of age, of which thirteen also at 6 months. Through a behavioural coding method and subsequent Principal Component Analysis (PCA), four reactivity components were identified. One stability component was identified, forming an operationalised definition of stability which is not yet well-defined. Behaviour was inconsistent within dogs between 6 and 13 months of age, suggesting that reactivity–stability expression at 6 months of age cannot predict expression at 13 months. A longitudinal approach is recommended to observe dogs from juvenile to adult life phases and assess when temperament emerges. Only one principal component, indicating sociability toward the handler, was associated with qualification. Overall, reactivity–stability had poor predictive validity for selection outcomes, undermining previous studies. Origin did predict outcomes: purchased dogs were less likely to be qualified or to be assigned the Assistance role than dogs from internal breeding lines. To identify factors affecting selection outcomes beyond the behavioural assessment, a questionnaire with puppy raisers is recommended. Investigating how reactivity–stability develops in the first year of life will enhance the assistance dog selection process.

Keywords: assistance dogs, temperament, reactivity–stability, consistency, validity

Introduction

Behavioural patterns differ between individuals and are often referred to as ‘temperament’ when consistent across time and situations (Réale et al., 2007). This phenomenon has also been termed ‘personality’ in human and animal behaviour (Carere & Maestripieri, 2013). Temperament tends to be consistent within one life stage but often shifts after a major developmental change (such as sexual maturation or metamorphosis) or stressful life events (Cabrera et al., 2021). Temperament is often described through dimensions. A common framework consists of five dimensions: exploration–avoidance, activity, shyness–boldness, sociability, or aggression (Réale et al., 2007), which have been observed mostly in mammals, as well as birds, reptiles, fish, and invertebrates (systematically reviewed by MacKinlay & Shaw, 2023). However, there is no agreed-upon definitive set of possible temperament dimensions, and terminology for the same behavioural patterns may not be aligned (Carter et al., 2013).

An animal’s temperament can be identified by observing its behaviour. Assessing an animal’s temperament by observing its behaviour can be done using observational and/or experimental approaches (Carter et al., 2013). In an observational approach, the animal is in its natural environment, whereas an experimental approach involves test batteries with different situations, tasks, and stimuli in a controlled environment. Importantly, temporal consistency is a prerequisite for behaviour to be considered a measure of temperament, thus behavioural assessments are often repeated at different time points to evaluate whether temperament can be identified (Bell et al., 2009). To gain a long-term overview of behaviour, a questionnaire approach may be used, which is collected from animal caretakers and pet owners who have spent a long time with an animal (Meagher, 2009). From their extensive knowledge of the animal’s temperament and habits, they rate the animal’s behaviour on items related to different temperament dimensions. This is often applied in animal husbandry settings.

Identifying temperament is an important practice in assistance dog selection (Bogaerts et al., 2019; Jones & Gosling, 2005). Assistance dogs (*Canis lupus familiaris*) are trained to aid people with disabilities in their daily lives and to participate in human society. In 2023, more than 1800 Dutch citizens lived with an assistance dog (European Guide Dog Federation, 2023). There are different roles for assistance dogs for different types of disabilities, including physical, medical, developmental, and psychiatric disabilities. Assistance dog programmes aim to select and train candidate dogs that naturally exhibit desirable temperaments for assistance work, to achieve high standards of quality, animal welfare, and efficiency. Having the right temperament is paramount to successful training and a functional ‘colleague’ relationship with the client (Bray et al., 2021). From an animal welfare perspective, facilitating the expression of natural behaviour is thought to be inherently rewarding, and is preferred over teaching dogs to perform desired but unnatural behaviour (Duranton & Horowitz, 2019). Releasing unsuitable candidates from the programme in a timely manner prevents negative experiences for the dog (Salmon et al., 2022).

Every year, up to 50% of candidates are disqualified from assistance dog programmes (in Belgium and the United States), of which behavioural issues are the second biggest cause, behind health issues (Bogaerts et al., 2019; Duffy & Serpell, 2012). Assistance dog programmes have often already invested time and financial resources in unsuitable candidates before the decision is made to release them. To save resources, assistance dog programmes conduct behavioural assessments with dogs as young as possible, usually within the first 18 months of life (Bray et al., 2021). However, this presents a biological problem. It may not be possible to accurately determine adult temperament while dogs are still in the juvenile phase (Serpell et al., 2016). A meta-analysis of 31 studies concluded that the canine temperament has significantly higher temporal consistency above 12 months of age than below 12 months (Fratkin et al., 2013). Temporal consistency, also called test-retest reliability or repeatability, is required for a behavioural pattern to be identifiable as a temperament dimension. Hence, the consensus is that temperament may not emerge until dogs reach adulthood.

One temperament dimension that may be promising in this regard is reactivity–stability. It is defined as the response to a novel, unexpected stimulus. An exemplary assistance dog shows stability, whereas reactivity is undesirable (Bogaerts et al., 2019; Bray et al., 2021). Behavioural definitions vary throughout literature. Stability, also referred to as resilience, is characterised by a relaxed body posture, high tail, showing no changes in behaviour after exposure to the stimulus, or approaching and exploring the stimulus without backing away, or recovering quickly from a startle response (Jones & Gosling, 2005; Lazarowski et al., 2021; McMahon et al., 2022). Stability is often defined, however, by a lack of reactivity behaviours. Reactivity is characterised by a startle response to the stimulus with a recovery period of more than one minute (Jones & Gosling, 2005). Asher et al. (2013) make a distinction between two forms of reactivity based on different coping styles: one a lack of confidence and the other overconfidence. Lack of confidence reactivity may be indicated by signs of fear, such as a tense or shaking body posture, low tail; or signs of avoidance, such as freezing, fleeing, or slowly backing up, or signs of stress such as lip licking and yawning (Dollion et al., 2019; Grigg et al., 2021; Harvey et al., 2016; Lazarowski et al., 2021). Overconfident reactivity may be indicated by signs of excitability, such as rapid tail wagging and barking toward stimuli, and intermittent avoidance and approach of novel stimuli (Grigg et al., 2021; Riemer et al., 2014), or signs of aggression such as growling and showing teeth (King et al., 2003). To evaluate reactivity–stability in candidate assistance dogs, behavioural assessments include tests that are designed to elicit reactivity, through stimuli mimicking regular occurrences in the daily life of an assistance dog (Lazarowski et al., 2021). Common tests are an acoustic startle (a sudden, loud noise is heard), novel object startle (an object suddenly appears close to the dog), and moving object startle (an object makes quick and unpredictable movements) (King et al., 2003).

There is tentative evidence that reactivity behaviours are consistent between juvenile and adult phases of the dog's life, and that reactivity can predict qualification and disqualification from assistance dog training and even the role for which dogs will be trained. Dollion et al. (2019) combined behavioural assessments with questionnaires with puppy raisers, and found that high reactivity at 6 and 12 months of age predicted disqualification. They also found that low reactivity predicted a guide dog role. Their definition of high reactivity was: excitability, pacing and running around, stimulus avoidance, barking, shaking (from Fratkin et al., 2013). Using similar methods, Harvey et al. (2016) and Lazarowski et al. (2021) found that low reactivity predicted qualification at 3, 8, and 11 months of age, though not at 5 months of age. They defined low reactivity as minor reactions, quick recovery, and not approaching the stimulus. Reactivity was consistent in the same dogs between 5 and 8 months of age (Harvey et al., 2016) and between 6 and 12 months of age (Dollion et al., 2019). The findings suggest that testing dogs on reactivity–stability below 12 months could give reliable insight into future behaviour at 12 months.

There are, however, important shortcomings to the research on reactivity–stability. Studies use different protocols to evaluate reactivity–stability and its predictive validity and temporal consistency. Various behavioural assessments, questionnaires, and physiological measurement tools exist and are often combined, which makes comparison of results difficult (Marcato et al., 2022). There is a need to validate findings based on different methods. Secondly, stability is often not defined beyond ‘the opposite of reactivity’. There is a need for knowledge of which behaviours indicate stability under novel, unexpected stimuli and situations, and whether these behaviours are indeed associated with qualification for assistance dog training (Rooney et al., 2016).

For this study, a behavioural assessment for the selection of assistance dog trainees was investigated. Hulphond Nederland is a non-profit organisation based in the Netherlands, testing more than 120 new candidate assistance dogs every year. Dogs are sourced from the programme’s breeding line or from external breeders. At 8 weeks of age, puppies are moved to a foster home and remain in their care until approximately 16 months. Candidate dogs are selected for training based on health indicators and performance during a behavioural assessment named CARAT (Clothier, 2023). Since CARAT has only been in use for five years, Hulphond Nederland are investigating how performance during the assessment relates to possible outcomes for each candidate. In an indoors open-field test battery, a dog handler guides the dog through several tests, and observers give a rating to the dog’s performance. The majority of candidate dogs take the assessment at 6 and 13 months of age. After, candidates may be qualified for training, or they may be disqualified and released form the programme. Qualified candidates are assigned one of three roles: (1) Assistance: mobility assistance, psychiatric service, and seizure response. The dogs have an assistance dog certification. (2) Therapy: animal-assisted therapy. (3) Breeding: The dogs take part in the internal breeding programme at Hulphond Nederland.

The aims of this study were to identify a reactivity–stability dimension in the CARAT behavioural assessment, to test whether this dimension changes between 6 and 13 months of age (temporal consistency), and to test whether it can predict qualification/disqualification and role (predictive validity). To achieve these aims, I observed dogs during an acoustic startle, novel object startle, and moving objects startle test. I expected that the reactivity–stability dimension could be identified. Temporal consistency was assessed through correlation tests on PCA-extracted component scores between the 6- and 13-month timepoints. I hypothesised that they would be temporally inconsistent, due to developing temperament in the first year of life. I assessed predictive validity by (1) comparing mean principal component scores between qualified and disqualified dogs and different roles, and (2) predictive modelling of qualification and role based on component scores. I hypothesised that stability would predict qualification and Assistance and Breeding roles, and that reactivity would predict disqualification and a Therapy role. These expectations were formed on the basis that stability is considered desirable whilst reactivity is undesirable, and Breeding dogs likely meet the highest behavioural standards, and Assistance dogs must achieve certification whereas Therapy dogs do not. Finally, I assessed predictive validity of individual characteristics – breed, sex, and origin – through statistically testing proportions of dogs with different characteristics and selection outcomes.

Methods

Subjects

Hulphond Nederland provided data of dogs that took the CARAT assessment between 2020 and February 2025. Of each subject was known: name, date of birth, sex, breed, origin (bred by the programme or purchased from external breeders), and assessment outcome (qualified or disqualified). If a dog was qualified, its role (Assistance, Therapy, or Breeding) was also known. For this study, I selected qualified subjects, and disqualified subjects that were disqualified based on behavioural reasons; dogs disqualified due to medical reasons or unknown reasons were excluded. Thirty-four subjects took the CARAT assessment at 13 months of age, of which thirteen subjects were also tested at 6 months of age. 27 were qualified (n=16 Assistance, n=8 Therapy, n=3 Breeding) and 7 were disqualified. The sex ratio was 18 female : 16 male. As for origin, 27 subjects came from in-programme breeding lines, and 20 from external breeders. As for breeds, there were 17 Labradors, 4 Golden Retrievers, 4 Labrador x Poodles, 3 Poodles, 3 Labrador x other crossbreeds, 2 undefined crossbreeds, and 1 Papillon (Appendix E).

Observations

The CARAT behavioural assessment took place in the same indoors test area at the Hulphond Nederland training facility. The area was fenced all around, with observers sitting at opposite ends behind the fence (**Figure 1**).

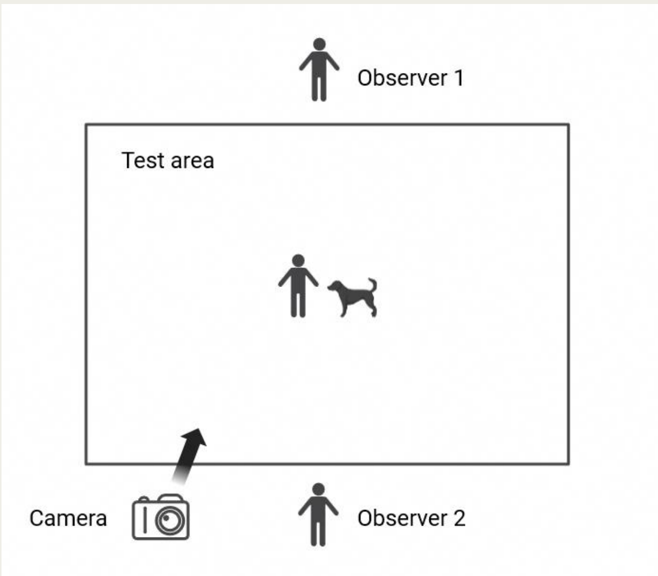


Figure 1. The CARAT behavioural assessment is conducted inside a fenced indoors area, with the handler and dog inside, and observers outside at opposite ends. Videos were taken from a corner angle and captured the entire test area.

The assessment consists of 15 tests designed to elicit behavioural responses to a range of situations and stimuli that resemble those commonly encountered by assistance dogs in the field. An overview of all tests can be found under Appendix A. Three tests named ‘Noise’, ‘Umbrella’, and ‘Car’ were observed. The assessment for 6-month-old dogs includes only ‘Noise’, whilst the assessment for 13-month-old dogs includes all three tests. Detailed descriptions of the tests of interest are as follows:

- Acoustic startle ‘Noise’: The handler distracts the dog by interacting with it in one corner of the test area. Suddenly, an observer on the opposite end of the test area drops a metal box filled with marbles. This creates a loud noise. The handler gives the dog 30 s to react, before walking up to the box. The handler shakes the box up and down three times. After waiting again for 30 s, the handler puts away the box.
- Novel object startle ‘Umbrella’: The handler guides the dog to a fixed spot and waits there. An observer holds an unopened umbrella a few metres opposite from the dog and at the dog’s eye level. Once the dog looks at the umbrella, it is opened in a sudden movement. The handler waits for the dog to react before walking to and interacting with the umbrella, to invite the dog to also interact with the object. The umbrella is put away once the dog has interacted or directed its attention elsewhere.
- Moving object startle ‘Car’: The handler shows a remote-controlled toy car to the dog, offering the opportunity to interact with it, before putting it on the floor. The car is moved three times from one end to the other on the field, pausing in-between.

Behaviour was observed with a coding method. An ethogram was constructed for reactivity–stability (Appendix B). Behaviours included in the ethogram were sourced from previous studies in which reactivity was observed in an acoustic startle test (Asher et al., 2013; Dollion et al., 2019; Marcato et al., 2024; Sforzini et al., 2009), videos of pet dogs’ responses to sudden loud noise in their home environment (Grigg et al., 2021), a novel object startle test (King et al., 2003; Riemer et al., 2014), and a moving object startle test (Harvey et al., 2016; Marcato et al., 2024). Four different forms of reactivity were identified from previous studies, which were included in the ethogram: stimulus avoidance, aggression, excitability, fearfulness. To contrast with stimulus avoidance, stimulus engagement behaviours were added. Body posture behaviours were added to distinguish between tense and relaxed body posture. Locomotion behaviours were also added.

Videos of the behavioural assessments were provided by Hulphond Nederland. Observations started 0.5 second before the stimulus was introduced and ended 0.5 second after the object was removed from the test area. Observations were recorded in BORIS, an open-source behavioural observation software (Friard & Gamba, 2016). The frequency and duration of behaviours in the ethogram was recorded. To prevent observer bias, the videos were viewed without knowledge of the subjects’ details or assessment outcomes.

Statistical analysis

RStudio (R Core Team, 2025) was used for statistical calculations. Significance was set at $P < 0.05$. The script is attached as supplementary material to this report. ChatGPT (OpenAI, 2025) was used to write parts of the R script. Please see the AI Usage Statement for details.

Extraction of reactivity–stability

To identify different forms of expression of reactivity–stability during the CARAT assessment, Principal Component Analysis (PCA) was used. Observations were compiled and normalised to reduce point event and state event data (i.e. scores and time intervals) to a similar scale. Three behaviours (*Growling, Whining, Showing_teeth*) were observed very few times or not at all. Hence, they were excluded from the further analysis. The covariance matrix of this dataset is attached to this report as supplementary material. A Varimax rotation was applied according to guidelines from Budaev (2010) for PCA with small sample sizes. The number of components to extract was chosen according to visual assessment of (average) eigenvalues in a scree plot, and computational assessment using parallel analysis. Parallel analysis produces a suggestion for the optimal number of components. PCA sorted groups of variables (i.e. behaviours) into components capturing the behavioural variation in the dataset – specifically, variation in the expression of a reactivity–stability dimension. Loadings were extracted, and loading values larger than 0.4 were used to interpret the components. When multiple components contained loadings >0.4 for the same behaviour, the behaviour was assigned to the component with the highest loading value.

Some limitations must be taken into account. The criteria for PCA with sample sizes smaller than 100 (in this study, n=34) include that behavioural measures must be tested for test-retest reliability and inter-rater reliability (Budaev, 2010). This could not be tested in the current study, but reliability was tested in the literature that was used to compile the ethogram. Another criterium is that loadings must have communalities larger than 0.25, ideally larger than 0.6 (Zeynivandnezhad et al., 2019), to meaningfully interpret PCA output. Communality for one behavioural measure is the sum of squared loadings on each principal component (McGarigal et al., 2000). Finally, another criterium is that only loadings larger than 0.5 should be selected for interpretation. In this study, this criterium was lowered to 0.4 to include more behavioural measures which allowed for more well-defined distinctions between the extracted components.

Temporal consistency of reactivity–stability

Change in reactivity–stability expression during the first year of life was investigated by correlating principal component scores of subjects that took the CARAT assessment twice, at the 6- and 13-month time points. First, a Shapiro-Wilk test for normality of component scores was performed, using residuals of component scores in a mixed-effects model. The model formula was as follows:

$$Component\ score \sim Age\ +\ (1\,|\,DogID)$$

The model predicted component score as a function of Age (fixed effect), with DogID (random effect) to account for non-independent observations from the same subjects. In case of normally distributed component scores, a Pearson correlation test was used, and in case of non-normally distributed component scores, a Spearman correlation test was used. Component scores at 13 months were condensed into a mean value to allow for correlation between two values, one for each timepoint. Change in reactivity–stability expression was tested both at the individual level (component score at 6 months on the x-axis and mean component score at 13 months on the y-axis) and at the group level (component scores on the y-axis and age on the x-axis).

Predictive validity of reactivity–stability on assessment outcomes

Associations between reactivity–stability and different assessment outcomes were investigated through comparing mean component scores between dogs with different selection outcomes, and through predictive modelling of outcomes. First, one-way ANOVAs were performed, testing whether each mean component score differed between qualified and disqualified dogs, and between different roles (Assistance, Therapy, and Breeding dogs). Post-hoc pairwise t-tests were done with a Bonferroni correction, due to unequal sample sizes.

Second, mixed-effect models were constructed to predict qualification (1) and role (2) according to component scores. Principal components are non-collinear and thus suitable to use as predictors. To verify this, VIFs of the fixed effect variables (i.e. component scores) were calculated for each model. If VIF > 5, the variable would be excluded. To account for non-independent observations of the same subjects, an identifier for individual subjects (DogID) was included as a random effect variable. Mixed-effect models require all random effects to be normally distributed; thus, the normality of DogID was checked with a Q-Q plot. Observations from 6-month-old subjects were left out as this would be a second source of non-independence.

Model (1) predicted the probability of qualification and disqualification as an assistance dog trainee. A logistic regression Generalised Linear Mixed Model (GLMM) was created, using a binomial distribution for binary dependent variables and a logit link function. A statistical power analysis was conducted on the GLMM, following guidelines from Kumle et al. (2021) for estimating power of mixed-effects models.

$$Outcome \sim PC1\ +\ PC2\ +\ PC3\ +\ PC4\ +\ PC5\ +\ (1\,|\,DogID)$$

Model (2) predicted the probability of being assigned to one of three roles: Assistance, Therapy, and Breeding (unordered levels). A Multinomial Logistic Regression Mixed Model (MLRMM) was constructed following directions from Gudmestad and Metzger (2025) for small sample sizes and incorporating mixed effects.

$$Role \sim PC1\ +\ PC2\ +\ PC3\ +\ PC4\ +\ PC5\ +\ (1\,|\,DogID)$$

For both models, odds ratios were calculated from variable coefficients. This indicates how the odds of qualification and different roles would change with every 1-point increase on the component scores. The significance of each variable was assessed with Type III Wald tests for the GLMM, and z-tests for the MLRMM. The null hypothesis is that a variable has no effect on the model, given that other variables in the model are kept constant. Goodness of fit was assessed with confusion matrices, McFadden's R-squared, and a Likelihood Ratio Test between the full model and a reduced model containing only the intercept and random effect. The null hypothesis is that the full model does not significantly improve goodness of fit compared to the reduced model.

Predictive validity of sample characteristics

Control variables (breed, sex, origin) were not included as model predictors because of collinearity and separation between these variables, causing convergence issues with logistic regression. Therefore, instead of predicting outcomes, proportions of different assessment outcomes across breed, sex, and origin groups were tested using Fisher's exact tests. Odds ratios were calculated indicating the odds of qualification and odds of assignment to a specific role. Breed was divided into four similarly sized groups: Golden Retriever, Labrador, Labrador mix, and Other (including all breed groups with less than 10 subjects).

Results

Extraction of reactivity–stability

PCA-extracted components had loadings with communalities between 0.08 - 0.92, mean 0.42±0.25. Communalities were deemed large enough to meaningfully interpret components. Five components were extracted, with 50.1% cumulative variance explained. The components were interpreted according to loadings larger than 0.4 (**Figure 2**).

The first component (PC1) contained 15.7% of the total variance and was interpreted as excitable stimulus engagement. It was loaded by *Running*, attentive body posture, stimulus gazing, independent stimulus approach without latency time, and interaction, and *Lunging* and *Barking*. Dogs with high scores in this component immediately and enthusiastically engaged with the stimulus. The second component (PC2) contained 12.0% of the total variance and was interpreted as fearful avoidance. It was loaded by *Walking*, low tail and tucked ears, *Panting*, stimulus gazing and keeping distance to the stimulus, *Barking*, and tail wagging. Dogs with high scores in this component were fearful of and did not engage with the stimulus. The third component (PC3) contained 8.2% of the total variance and was interpreted as independent avoidance. It was positively loaded by keeping distance to both handler and stimulus and gazing elsewhere, *Fleeing* and *Freezing*, and a long latency time (<30 s) before approaching the stimulus independently. It was negatively loaded by staying close to the handler. Dogs with high scores on this component avoided both the stimulus and handler, and were hesitant to approach the stimulus. Dogs with low scores sought out the handler. The fourth component (PC4) contained 8.0% of the total variance and was interpreted as fearful handler-oriented focus. It was loaded by *Standing*, gazing and proximity to the handler, handler-guided stimulus approach after a long latency time (>30 s), fearful body posture, *Trembling*. Dogs with high scores on this component were fearful but followed the handler to the stimulus.

The fifth component (PC5) contained 6.1% of the total variance and was interpreted as fearfulness or quick recovery from startle. It was positively loaded by low tail and *Panting*. It was negatively loaded by *Flinching,* attentive body posture, handler-guided stimulus approach without latency time. Dogs with high scores on this component displayed fearfulness. Dogs with low scores on this component had a startle response, but quickly recovered and engaged with the stimulus.

Figure 2. Loading matrix of the PCA on behavioural measures. For interpretation of the components, loading values larger than 0.4 were selected. They are in bold and marked as positive (white) or negative (red). PC1 to PC4 indicated different forms of reactivity, and PC5 indicated stability.

Behaviour	PC1	PC2	PC3	PC4	PC5
Approach_stimulus_handler_fast	-0.024	0.203	-0.275	-0.089	**-0.589**
Approach_stimulus_handler_slow	-0.091	-0.080	-0.097	**0.469**	0.068
Approach_stimulus_indep_fast	**0.814**	-0.016	0.055	-0.219	0.279
Approach_stimulus_indep_slow	0.037	-0.037	**0.510**	-0.069	0.079
Approach_stimulus_not	-0.189	**0.557**	0.266	0.096	-0.256
Backing_up	0.048	**0.720**	-0.095	-0.169	-0.163
Barking	**0.457**	**0.471**	0.012	-0.112	-0.341
Fleeing	-0.061	0.048	**0.526**	0.090	-0.014
Flinching	-0.385	0.042	0.108	-0.048	**-0.474**
Freezing	-0.161	0.087	**0.505**	0.182	0.136
Lunging	**0.824**	0.080	0.056	-0.145	0.102
Orientation_elsewhere	0.149	0.212	**0.433**	0.071	-0.024
Orientation_handler	0.183	0.126	-0.103	**0.745**	-0.134
Orientation_stimulus	**0.547**	**0.578**	0.300	0.061	0.049
Panting	0.091	**0.608**	-0.029	-0.118	**0.475**
Proximity_handler_close	0.032	0.340	**-0.605**	**0.547**	-0.160
Proximity_handler_far	**0.520**	0.359	**0.708**	-0.008	0.149
Proximity_stimulus_close	**0.886**	-0.133	-0.238	0.045	-0.052
Proximity_stimulus_far	-0.051	**0.747**	**0.440**	0.352	0.055
Running	**0.765**	0.114	0.306	-0.111	0.016
Standing	0.382	0.397	0.124	**0.590**	-0.082
Stimulus_interaction	**0.754**	-0.277	-0.130	-0.060	0.110
Tail_low	-0.248	0.334	0.215	**0.529**	**0.400**
Tail_low_medium	-0.152	**0.531**	-0.182	0.104	**0.490**
Tail_medium_high	**0.712**	0.139	0.257	-0.001	**-0.521**
Trembling	-0.077	0.015	-0.046	**0.471**	0.128
Tucked_ears	-0.334	**0.501**	-0.040	-0.055	0.089
Wagging_relaxed_excited	**0.639**	**0.503**	-0.352	-0.004	-0.080
Wagging_tense	0.006	-0.115	0.220	**0.578**	-0.236
Walking	0.218	**0.717**	0.136	0.176	0.048

Temporal consistency of reactivity–stability

Between 6 and 13 months, reactivity–stability expression was consistent at the group level (**Table 1A**). Contrastingly, at the individual level, reactivity–stability expression varied with age (**Table 1B**). Graphical representations of the correlation tests are included under Appendix C. It should be noted that, despite not reaching significance, PC2 (fearful avoidance) and PC5 (quick recovery from startle) had moderately high correlation coefficients ($r > 0.3$), indicating that the group had higher scores on these components at 13 months than at 6 months, but individuals were more consistent in these components than others.

Table 1A. At the group level, reactivity–stability was consistent between 6 and 13 months of age. Principal component scores were tested for correlations over time. No significant correlations were found. Abbreviations: Normality: results of Shapiro-Wilk tests for normality on component scores. P: P-values. Correlation test: non-normally distributed component scores were tested with a Spearman correlation, and normally distributed component scores were tested with a Pearson correlation. r: correlation coefficient.

Component	Shapiro-Wilks P	Normality	Correlation test	r	r P
PC1	< 0.01	×	Spearman	0.126	0.683
PC2	< 0.01	×	Spearman	0.440	0.135
PC3	< 0.01	×	Spearman	0.055	0.863
PC4	< 0.01	×	Spearman	-0.220	0.470
PC5	0.165	✓	Pearson	0.302	0.316

Table 1B. At the individual level, reactivity–stability was inconsistent between 6 and 13 months of age. Principal component scores of individual dogs at 6 months were tested for correlations with scores at 13 months. No significant correlations were found between component scores.

Component	Shapiro-Wilks P	Normality	Correlation test	r	r P
PC1	0.027	×	Spearman	0.126	0.683
PC2	0.694	✓	Pearson	0.311	0.301
PC3	0.013	×	Spearman	0.055	0.863
PC4	0.741	✓	Pearson	-0.046	0.882
PC5	0.470	✓	Pearson	0.302	0.316

Predictive validity of reactivity–stability on assessment outcomes

Only PC3 (independent avoidance) had significantly different mean principal component scores between qualified and disqualified dogs (ANOVA: $F_{1,95} = 5.336$, $P = 0.023$). Qualified dogs had a significantly lower mean score on PC3 than disqualified dogs (disqualified - qualified = 0.580±0.251, paired t-test: $t_{95} = 2.310$, $P = 0.023$). This indicated that qualified dogs on average stayed near the handler. A graphical representation of the pairwise comparisons between qualified and disqualified dogs for each principal component is included under Appendix D. There were no significant differences in mean component scores between different roles (Assistance, Therapy, and Breeding) (Ps > 0.1).

Logistic regression models failed to significantly predict probabilities of selection outcomes of candidate dogs at 13 months of age. The GLMM, predicting qualification for assistance dog training, had poor goodness of fit (R-squared = 0) and was not more accurate than a null model without variables (LRT: $\chi^2_5 = 0.114$, $P = 1$). The estimated power of the GLMM was 0.10. According to a confusion matrix, the model was unable to make true predictions. Variable coefficients failed to reach significance according to Type III Wald tests, except for the intercept (**Table 2A**). This indicated that regardless of component scores, dogs were significantly likely to be qualified. The MLRMM, predicting qualification for Assistance over Breeding and Therapy over Breeding, had moderate goodness of fit (R-Squared = 0.523) but was not significantly better than a null model (LRT: $\chi^2_{10} = 4.231$, $P = 0.936$). According to the confusion matrix, the model was able to make true predictions for all qualified subjects. Variable coefficients failed to reach significance according to z- tests, except for the intercept of Assistance over Breeding (**Table 2B**). This indicated that regardless of principal component scores, Assistance dogs were significantly more likely to be assigned to Assistance than to Breeding.

Table 2A. Reactivity–stability could not predict qualification as an assistance dog trainee. A logistic regression GLMM was constructed to predict qualification based on principal component scores. Abbreviations: OR: odds ratios of probability of qualification over disqualification. P: P-values of Chi-Squared values from Type III Wald tests on coefficients.

Variable	Coefficient	OR	% Change in odds	P
Intercept	13.604			< 0.01
PC1	-0.056	0.946	-5.40%	0.978
PC2	0.502	1.652	+65.20%	0.854
PC3	-0.646	0.524	-47.60%	0.765
PC4	0.381	1.464	+46.40%	0.892
PC5	0.246	1.278	+27.80%	0.908

Table 2B. Reactivity–stability could not predict assignment to Assistance over Breeding or Therapy over Breeding. A MLRMM was constructed to predict assignment to Assistance over Breeding and Therapy over Breeding. Abbreviations: OR: odds ratios of probability of being assigned to the Assistance or Therapy role over Breeding. P: P-values of z-tests on coefficients.

Variable	Coefficient	OR	% Change in odds	P
Assistance ~Intercept	2.119			< 0.01
Therapy ~Intercept	0.878			0.267
Assistance ~PC1	0.359	1.433	+43.30%	0.704
Therapy ~PC1	0.272	1.313	+31.30%	0.777
Assistance ~PC2	0.742	2.100	+110.00%	0.412
Therapy ~PC2	1.098	3.000	+200.00%	0.225
Assistance ~PC3	-0.604	0.547	-45.30%	0.428
Therapy ~PC3	0.007	1.007	+0.70%	0.992
Assistance ~PC4	0.157	1.170	+17.00%	0.851
Therapy ~PC4	-0.057	0.945	-5.50%	0.949
Assistance ~PC5	0.278	1.320	+32.00%	0.701
Therapy~PC5	-0.033	0.967	-3.30%	0.963

Predictive validity of individual characteristics on assessment outcomes

Associations were tested between individual characteristics (origin, breed, and sex) and qualification/disqualification, as well as Assistance, Therapy, and Breeding. Purchased dogs had significantly lower odds of qualification than dogs from in-house breeding lines at Hulphond Nederland (OR = 0.136, CI = 0.0143 – 0.630, P = 0.004). Purchased dogs also had significantly lower odds of qualifying for the Assistance role than dogs from in-house breeding lines (OR = 0.351, CI = 0.115 – 1.003, P = 0.037). No other significant associations between roles an origin were found (Ps > 0.2). As for breed, Labradors had significantly lower odds of qualification than other breed groups (OR = 0.267, CI = 0.0689 – 0.872, P = 0.023). No other significant associations between breed groups and qualification were found (Ps > 0.6). All breed groups were significantly more likely to be assigned to Assistance than other roles (P = 0.041). As for sex, there was no significant association with qualification (P > 0.1). However, females were significantly more likely to be assigned to Breeding, because exclusively female subjects got this role (P = 0.002).

Discussion

Extraction of reactivity–stability

The first aim of this study was to identify a reactivity–stability dimension in 6- and 13-month-old dogs during the CARAT behavioural assessment. In accordance with my expectation, reactivity–stability behaviours were observed in all subjects. PC1 (excitable stimulus engagement), PC2 (fearful avoidance), PC3 (independent avoidance), PC4 (fearful handler-oriented focus), and positive loadings on PC5 (fearfulness) were interpreted as different forms of reactivity. Negative loadings on PC5 (quick recovery from startle) were interpreted as stability. The findings confirm that scoring behaviours during a behavioural assessment including acoustic startle, novel object startle, and moving object startle tests can effectively evaluate reactivity–stability in all its different forms.

The results also illustrate a problem: reactivity–stability takes different forms, and as a result, it is difficult to distinguish from other dimensions, and its definition is not aligned across studies. Fearfulness was found in several components: PC2 (fearful avoidance), PC4 (fearful handler-oriented focus), and positive loadings on PC5 (fearfulness). Indeed, a commonly used model of canine temperament developed by Jones and Gosling (2005), after meta-analysis of 51 studies and sorting behaviours into categories done by experts, saw the most overlap between reactivity and fearfulness. The study concluded that these two categories are distinguished as follows: reactivity is the opposite of stability, and fearfulness is the opposite of confidence. In more recent research, reactivity is sometimes conflated with fearfulness (Fratkin et al., 2013), and sometimes seen as separate dimensions (McGarrity et al., 2015). My findings support the framework that reactivity takes different forms as it relates to the dog's coping style to novel stimuli. Dogs may be overconfident/active, or non-confident/passive (Asher et al., 2013). Following the findings of this study, fearfulness and avoidance are forms of non-confident reactivity whilst excitability is a form of over-confident reactivity. Contrary to the definition given by Sforzini et al. (2009), aggression (indicated by growling, showing teeth, barking) was not found to be a form of reactivity and thus these terms should not be conflated. Going forward, a solution to the problem of varying definitions for reactivity–stability may be to use operational definitions that precisely convey what studies are measuring, and to leave interpretations of behaviour as part of temperament for the discussion (Laskowski et al. (2022).

Moreover, the results provide a suggestion for an operationalised definition of stability. Negative loadings on PC5 (quick recovery from startle) were: flinching, handler-guided stimulus approach without latency time, and an upright tail position. Dogs would startle from the stimulus, but then quickly moved on to an attentive body posture and followed the handler's guidance toward the stimulus. This aligns with the definition of stability from Jones & Gosling (2005): high tail position, recovering quickly from a startle response to novel stimuli, and approaching and exploring novel stimuli without backing away. Critically, I did not interpret 'showing no changes in behaviour after exposure to the stimulus' as stability. My results suggest that stability cannot be defined as a non-reaction, but rather, that quick recovery from startle may be an indicator of successful coping to startling stimuli. Lensen et al. (2019) conducted a behavioural assessment including a moving object test and object startle test, combined with physiological stress marker measurements and a questionnaire with dog owners (Serpell & Hsu, 2001). They found that elevated stress indicators at the 10-minute mark after the behavioural assessment were associated with low fearfulness, according to owners' descriptions of the temperaments of their dogs. Dogs with lower initial stress levels needed 40 minutes to recover to baseline, and frequently showed fearfulness behaviours according to their owners. The study concluded that a sharp increase in stress levels followed by a rapid decrease back to baseline reflects an adaptive stress response to novel stimuli. A startle response by itself may not be an indicator of unsuitability to assistance work if it is accompanied by quick recovery. To enhance selection for stability, I recommend that future research explore the duration of time before candidate assistance dogs recover from a startle response, in relation to reactivity–stability and to selection outcomes.

A temperament dimension unrelated to reactivity–stability was also found: PC3 (independent avoidance) and PC4 (fearful handler-oriented focus) indicated a behavioural pattern that was not explicitly outlined in the ethogram. PC3 was negatively loaded by proximity to the handler, and PC4 was positively loaded by gazing and proximity to the handler. Handler focus and handler avoidance reflect whether dogs give and seek attention from humans or elsewhere. In the literature, this is categorised as a sociability–lack of interest in others dimension (Jones & Gosling, 2005). Sociable pet dogs have been found to gaze longer at unfamiliar humans than less sociable dogs (Jakovcevic et al., 2012).

Temporal consistency of reactivity–stability

The second aim of this study was to assess temporal consistency of reactivity–stability expression between 6 and 13 months of age. In line with the hypothesis, individual temporal consistency was not shown, indicating that the same dogs had varied reactivity–stability between 6 and 13 months of age. Seemingly contrary to the hypothesis, reactivity–stability expression was temporally consistent at the group level. The results indicate that reactivity–stability was consistent over time in some dogs, but varied in others. This aligns with the consensus in literature: during the first year of a dog's life, behavioural patterns often change, and temporal consistency is low (Fratkin et al., 2013). Though not significant, individual-level temporal consistency was moderately high for PC2 (fearful avoidance) and PC5 (fearfulness or quick recovery from startle). Specifically, dogs appeared to show fearful body posture (PC2 and PC5) and to keep distance to the stimulus (PC2) somewhat consistently. Consistent behaviours between juvenile and adult dogs may have a strong genetic component; fearfulness may be one such highly heritable dimension (MacLean et al., 2019). Further research could investigate the heritability of reactivity–stability, which overlaps with fearfulness.

It is of high interest of assistance dog training programmes to find the earliest age possible to do behavioural assessments that give reliable estimates of dogs' temperament in the future. The findings of this study imply that behavioural assessments done during at 6 months cannot offer insight into dogs' reactivity–stability expression at 13 months. Early life experiences, socialisation, sexual maturation, and genetics all shape temperament, and all factors could be explored in future research (Ilska et al., 2017). Innovative selection methods could include a broad background investigation, with a questionnaire for puppy raisers, to identify environmental factors that may have influenced temperament development. For example, if handlers or foster carers do not let a dog interact with strangers often, it is more likely to exhibit fearfulness or avoidance when in contact with humans than a well-socialised and positively reinforced dog (Rooney et al., 2016; Salonen et al., 2023). Furthermore, ontogeny of temperament may take up to 18-24 months, especially for larger dog breeds like Labrador Retrievers (Serpell et al., 2016). At Hulphond Nederland and other assistance dog programmes across the world, Labrador Retrievers are a very common dog breed. I recommend that future studies take a longitudinal approach, conducting behavioural measurements at multiple timepoints below and above the first 12 months of life. For example, testing at 6, 12, 18, 24 and 30 months will give insight into temperament development during and beyond the juvenile phase. This could also be a practical application for Hulphond Nederland. In combination with knowledge on the circumstances of ontogeny, this will improve understanding of reactivity–stability and other temperament dimensions.

Predictive validity of reactivity–stability on assessment outcomes

The third aim was to assess predictive validity of reactivity–stability at 13 months of age for different assessment outcomes. I hypothesised that low reactivity and high stability would be associated to a higher likelihood of qualification and of assignment to Breeding and Assistance roles, and that high reactivity would predict disqualification and the Therapy role.

Only PC3 (independent avoidance) had a significantly lower mean score in qualified dogs than disqualified dogs. Low (negative) scores on PC3 indicated close proximity to the handler, which is interpreted as sociability. Indeed, sociability is a desirable dimension in assistance dogs to be able to work together well with clients and trainers (Bray et al., 2021; Lazarowski et al., 2019). Contrastingly, no positive association between PC4 (fearful handler-oriented focus) and qualification was found. In this component, the combination of sociability with fearfulness, an undesirable dimension, likely lowered the likelihood of qualification. These findings can be applied to assistance dog selection: observers could screen for dogs that seek out the handler in a confident state and that appear sociable, rather than dogs that seek out the handler in a fearful or anxious state.

Contrary to my hypothesis, no other principal components were associated with selection outcomes, and none of the principal components could predict qualification/disqualification or role in logistic regression models. Overall, the results indicate that reactivity–stability expression at 13 months in a behavioural assessment has poor predictive validity for assistance dog qualification and role. The results negatively compare to the previous research finding that reactivity behaviours can accurately predict disqualification from assistance dog programmes when assessed using a combination of behavioural assessments, questionnaires, and physiological measurement (Bogaerts et al., 2019; Bray et al., 2021). Instead, all dogs were likely to qualify for assistance dog training, and to be assigned to the Assistance role, regardless of reactivity–stability.

Firstly, this result was likely due to methodology; in this study, a sample of 34 dogs was observed which was also skewed toward qualified dogs – 27 dogs were qualified, while only 7 were disqualified. However, the sample was sourced in an unbiased manner. A few dogs did not complete the assessment due to being too dysregulated or distracted to continue. While the decision to prematurely quit the assessment is done in accordance with animal welfare standards, consequently, this reduces the opportunities to observe the behaviour of disqualified subjects. Another possible methodological issue was the behavioural assessment design. Acoustic startle, novel object startle, and moving object startle tests may not be the only tests in which reactivity–stability may be observed. Studies that found good predictive validity of reactivity–stability for selection outcomes had a wider variation of tests and stimuli that may elicit a startle response, such as meeting a stranger or being touched like during a veterinary exam (Dollion et al., 2019; Harvey et al., 2016; Lazarowski et al., 2021). Future research could isolate such test to assess predictive validity of performance during each test for selection outcomes. Furthermore, test execution was not always strictly the same: for example, (not) waiting 30 seconds allowing the dog to explore the stimulus independently, or (not) waiting for the dog to sit and direct their attention forward before opening the umbrella. This may have affected dogs' performance and test validity.

Secondly, this result may have been due to factors other than dog behaviour during the behavioural assessment, that influenced selection outcomes but could not be captured in this study. Testing observer bias was outside of the scope of this study, because observer identity could not be gathered from the videos. However, it seems likely that their unfamiliarity with the dogs made it difficult to assess temperament on the basis of performance during the behavioural assessment alone. Future studies could take dog behaviour into account outside the context of the behavioural assessment through a subjective rating method. This takes the form of a questionnaire taken by puppy raisers, who assign scores to different behavioural patterns based on their experiences with the dog over a long period of time and a wide range of situations. An example of a subjective rating method is the C-BARQ (Serpell & Hsu, 2001), which is intended to capture problematic behaviours in pet dogs. It was found to have good predictive validity for guide dog and mobility assistance dog qualification in a population of 6- and 12-month-old subjects (Duffy & Serpell, 2012). Combining C-BARQ with behavioural assessments has been found to predict selection outcomes, specifically the likelihood of disqualification (Bray et al., 2019). It has become a standard tool in canine behaviour research. Though Hulphond Nederland may already consult puppy raisers in an informal way, I recommend adding a formalised questionnaire method to gather behavioural data outside the behavioural assessment. This will allow for more informed decisions on selection outcomes.

Predictive validity of individual characteristics on assessment outcomes

Finally, individual characteristics (sex, breed, origin) in the sample were investigated in relation to assessment outcomes. Dogs purchased from external breeders were significantly less likely to qualify and also less likely to be assigned an Assistance role than dogs from in-programme breeding lines. However, all breeds in the sample of this study were more likely to qualify for Assistance dog training than other roles. These results imply that an observer bias affected CARAT assessment outcomes in favour of dogs bred in-programme, regardless of purchased dogs' performance. This stands in contrast to evidence that dogs purchased from breeders have greater success as a guide dog than dogs bred in-programme (Batt et al., 2010). One solution to combat observer bias for origin may be to blind observers to this information.

As for breed, Labrador Retrievers were less likely to qualify than other breed groups. However, Labradors were by far the largest breed group in the sample, and disqualified subjects in other breed groups were sparse or non-existent causing data separation (all twelve Golden Retrievers qualified). Labrador Retrievers are a very common pet dog and assistance dog breed, and are associated with high working success as assistance dogs (Bremhorst et al., 2018). Therefore, the negative association with qualification in this study likely cannot be replicated with a larger sample.

Conclusion

Reactivity–stability expressed during a behavioural assessment was successfully evaluated, but previous research findings of temporal consistency and predictive validity for selection outcomes were not confirmed. A definition for stability as 'quick recovery from startle' was proposed, incorporating insights from behavioural observations and research on physiological signs of successful coping to reactivity-inducing stimuli. Behaviour during acoustic startle, moving object, and novel object tests with dogs at 6 months of age could not offer insight into the same dogs' behaviour at 13 months. Longitudinal studies ranging from 6- to 24 month-old timepoints are recommended to investigate when temperament emerges. Associations between the reactivity–stability dimension and selection outcomes in the Hulphond Nederland assistance dog programme were limited. Only low scores on PC3 (independent avoidance), indicating sociability rather than stability, predicted qualification as an assistance dog trainee. This is contrasted by PC4 (fearful handler-oriented focus), which did not have the same association with success. Regardless of reactivity–stability, many dogs were qualified. There may have been other factors that influenced selection outcomes outside of the behavioural assessment, which could be explored by collecting information about the dog's foster home environment. With more research on factors that affect development of the reactivity–stability dimension within the first year of life, the assistance dog selection process can be enhanced.

AI Usage Statement

ChatGPT, a generative AI application (OpenAI, 2025), was used to write functions and solve errors in the R script for statistical analysis. Prompts specified available variables, desired output, and format. No data was uploaded. After use of this application, the author reviewed and edited the content as needed and takes full responsibility for the contents of this report.

Acknowledgements

I would like to thank Annika van der Wal, our contact person at Hulphond Nederland, for very patiently and graciously answering our questions about the organisation and the training process for assistance dog candidates, and for sharing the (freshly updated!) dataset with me.

I would like to thank Jori Noordenbos, Marc Naguib, and Giulia Cimarelli for their guidance, poignant and invaluable points of feedback, for going out of their way to help me make sense of RStudio output, and for showing me the way around a research field that was very new to me at the start.

I would like to thank my neighbours at Haarweg 191-145 for their invaluable support, working together in our living room, gym buddies, and surprising me with freshly baked cookies on late-night work sessions.

I would like to thank my friends from the 2019 cohort of Biology, who are fantastic researchers in their own right, for reassuring me whenever I needed direction, or simply sharing a cosy dinner with me.

I would like to thank my parents for their faith in me, their enthusiasm in following my journey, and for encouraging me to see obstacles along the way as learning opportunities.

Finally, I would like to thank Soraya from WUR Writing Lab, who helped me take the first steps.

References

Asher, L., Blythe, S., Roberts, R., Toothill, L., Craigon, P. J., Evans, K. M., Green, M. J., & England, G. C. W. (2013). A standardized behavior test for potential guide dog puppies: Methods and association with subsequent success in guide dog training. *Journal of Veterinary Behavior*, *8*(6), 431-438. https://doi.org/10.1016/j.jveb.2013.08.004

Batt, L., Batt, M., Baguley, J., & McGreevy, P. (2010). Relationships between puppy management practices and reported measures of success in guide dog training. *Journal of Veterinary Behavior*, *5*(5), 240-246. https://doi.org/10.1016/j.jveb.2010.02.004

Bell, A. M., Hankison, S. J., & Laskowski, K. L. (2009). The repeatability of behaviour: a meta-analysis. *Anim Behav*, *77*(4), 771-783. https://doi.org/10.1016/j.anbehav.2008.12.022

Bogaerts, E., Moons, C., Van Nieuwerburgh, F., Peelman, L., Saunders, J., & Broeckx, B. (2019). Rejections in an non-purpose bred assistance dog population: Reasons, consequences and methods for screening. *PLoS ONE*, *14*. https://doi.org/10.1371/journal.pone.0218339

Bray, E. E., Levy, K. M., Kennedy, B. S., Duffy, D. L., Serpell, J. A., & MacLean, E. L. (2019). Predictive Models of Assistance Dog Training Outcomes Using the Canine Behavioral Assessment and Research Questionnaire and a Standardized Temperament Evaluation. *Front Vet Sci*, *6*, 49. https://doi.org/10.3389/fvets.2019.00049

Bray, E. E., Otto, C. M., Udell, M. A. R., Hall, N. J., Johnston, A. M., & MacLean, E. L. (2021). Enhancing the Selection and Performance of Working Dogs. *Front Vet Sci*, *8*, 644431. https://doi.org/10.3389/fvets.2021.644431

Bremhorst, A., Mongillo, P., Howell, T., & Marinelli, L. (2018). Spotlight on Assistance Dogs—Legislation, Welfare and Research. *Animals,* *8*(8), 129. https://doi.org/doi:10.3390/ani8080129

Budaev, S. V. (2010). Using Principal Components and Factor Analysis in Animal Behaviour Research: Caveats and Guidelines. *Ethology*, *116*(5), 472-480. https://doi.org/10.1111/j.1439-0310.2010.01758.x

Cabrera, D., Nilsson, J. R., & Griffen, B. D. (2021). The development of animal personality across ontogeny: a cross-species review. *Animal Behaviour*, *173*, 137-144. https://doi.org/10.1016/j.anbehav.2021.01.003

Carere, C., & Maestripieri, D. (2013). Animal Personalities: Behavior, Physiology, and Evolution. University of Chicago Press. https://doi.org/10.7208/chicago/9780226922065.001.0001

Carter, A. J., Feeney, W. E., Marshall, H. H., Cowlishaw, G., & Heinsohn, R. (2013). Animal personality: what are behavioural ecologists measuring? *Biological Reviews*, *88*(2), 465-475. https://doi.org/10.1111/brv.12007

Clothier, S. (2023, 3 April). *CARAT*. https://suzanneclothier.com/carat/

Dollion, N., Paulus, A., Champagne, N., St-Pierre, N., St-Pierre, É., Trudel, M., & Plusquellec, P. (2019). Fear/reactivity in working dogs: An analysis of 37 years of behavioural data from the Mira Foundation's future service dogs. *Applied Animal Behaviour Science*, *221*, 104864. https://doi.org/10.1016/j.applanim.2019.104864

Duffy, D. L., & Serpell, J. A. (2012). Predictive validity of a method for evaluating temperament in young guide and service dogs. *Applied Animal Behaviour Science*, *138*(1), 99-109. https://doi.org/10.1016/j.applanim.2012.02.011

Duranton, C., & Horowitz, A. (2019). Let me sniff! Nosework induces positive judgment bias in pet dogs. *Applied Animal Behaviour Science*, *211*, 61-66. https://doi.org/10.1016/j.applanim.2018.12.009

European Guide Dog Federation. (2023). *2023 Survey on Assistance Dogs in Europe* [PDF]. https://www.egdfed.org/media/1138/european-guide-dog-federation-2023-survey.pdf

Fratkin, J. L., Sinn, D. L., Patall, E. A., & Gosling, S. D. (2013). Personality consistency in dogs: a meta-analysis. *PLoS ONE*, *8*(1), e54907. https://doi.org/10.1371/journal.pone.0054907

Friard, O., & Gamba, M. (2016). BORIS: a free, versatile open‐source event‐logging software for video/audio coding and live observations. *Methods in ecology and evolution*, *7*(11), 1325-1330. https://doi.org/10.1111/2041-210X.12584

Grigg, E. K., Chou, J., Parker, E., Gatesy-Davis, A., Clarkson, S. T., & Hart, L. A. (2021). Stress-Related Behaviors in Companion Dogs Exposed to Common Household Noises, and Owners' Interpretations of Their Dogs' Behaviors. *Front Vet Sci*, *8*, 760845. https://doi.org/10.3389/fvets.2021.760845

Gudmestad, A., & Metzger, T. A. (2025). Mixed-Effects Modeling with a Multinomial Dependent Variable. *Language Learning*, *75*(1), 212-249. https://doi.org/10.1111/lang.12667

Harvey, N. D., Craigon, P. J., Sommerville, R., McMillan, C., Green, M., England, G. C. W., & Asher, L. (2016). Test-retest reliability and predictive validity of a juvenile guide dog behavior test. *Journal of Veterinary Behavior*, *11*, 65-76. https://doi.org/10.1016/j.jveb.2015.09.005

Ilska, J., Haskell, M. J., Blott, S. C., Sánchez-Molano, E., Polgar, Z., Lofgren, S. E., Clements, D. N., & Wiener, P. (2017). Genetic Characterization of Dog Personality Traits. *Genetics*, *206*(2), 1101-1111. https://doi.org/10.1534/genetics.116.192674

Jakovcevic, A., Mustaca, A., & Bentosela, M. (2012). Do more sociable dogs gaze longer to the human face than less sociable ones? *Behavioural Processes*, *90*(2), 217-222. https://doi.org/10.1016/j.beproc.2012.01.010

Jones, A. C., & Gosling, S. D. (2005). Temperament and personality in dogs (Canis familiaris): A review and evaluation of past research. *Applied Animal Behaviour Science*, *95*(1), 1-53. https://doi.org/10.1016/j.applanim.2005.04.008

King, T., Hemsworth, P. H., & Coleman, G. J. (2003). Fear of novel and startling stimuli in domestic dogs. *Applied Animal Behaviour Science*, *82*(1), 45-64. https://doi.org/10.1016/S0168-1591(03)00040-6

Kumle, L., Võ, M. L. H., & Draschkow, D. (2021). Estimating power in (generalized) linear mixed models: An open introduction and tutorial in R. *Behavior Research Methods*, *53*(6), 2528-2543. https://doi.org/10.3758/s13428-021-01546-0

Laskowski, K. L., Chang, C.-C., Sheehy, K., & Aguiñaga, J. (2022). Consistent Individual Behavioral Variation: What Do We Know and Where Are We Going? *Annual Review of Ecology, Evolution, and Systematics*, *53*(Volume 53, 2022), 161-182. https://doi.org/10.1146/annurev-ecolsys-102220-011451

Lazarowski, L., Rogers, B., Krichbaum, S., Haney, P., Smith, J. G., & Waggoner, P. (2021). Validation of a Behavior Test for Predicting Puppies' Suitability as Detection Dogs. *Animals*, *11*(4), 993. https://doi.org/10.3390/ani11040993

Lazarowski, L., Strassberg, L. R., Waggoner, L. P., & Katz, J. S. (2019). Persistence and human-directed behavior in detection dogs: Ontogenetic development and relationships to working dog success. *Applied Animal Behaviour Science*, *220*, 104860. https://doi.org/10.1016/j.applanim.2019.104860

Lensen, R. C. M. M., Moons, C. P. H., & Diederich, C. (2019). Physiological stress reactivity and recovery related to behavioral traits in dogs (Canis familiaris). *PLoS ONE*, *14*(9), e0222581. https://doi.org/10.1371/journal.pone.0222581

Ley, J., Bennett, P., & Coleman, G. (2008). Personality dimensions that emerge in companion canines. *Applied Animal Behaviour Science*, *110*(3), 305-317. https://doi.org/10.1016/j.applanim.2007.04.016

MacKinlay, R. D., & Shaw, R. C. (2023). A systematic review of animal personality in conservation science. *Conservation Biology*, *37*(1), e13935. https://doi.org/10.1111/cobi.13935

MacLean, E. L., Snyder-Mackler, N., vonHoldt, B. M., & Serpell, J. A. (2019). Highly heritable and functionally relevant breed differences in dog behaviour. *Proceedings of the Royal Society B: Biological Sciences*, *286*(1912), 20190716. https://doi.org/10.1098/rspb.2019.0716

Marcato, M., Kenny, J., O'Riordan, R., O'Mahony, C., O'Flynn, B., & Galvin, P. (2022). Assistance dog selection and performance assessment methods using behavioural and physiological tools and devices. *Applied Animal Behaviour Science*, *254*, 105691. https://doi.org/10.1016/j.applanim.2022.105691

Marcato, M., Tedesco, S., O'Mahony, C., O'Flynn, B., & Galvin, P. (2024). Prediction of working outcomes in trainee dogs using the novel Assistance Dog Test Battery (ADTB). *Applied Animal Behaviour Science*, *272*, 106212. https://doi.org/10.1016/j.applanim.2024.106212

McGarigal, K., Cushman, S., & Stafford, S. G. (2000). *Multivariate statistics for wildlife and ecology research*. Springer. https://doi.org/10.1007/978-1-4612-1288-1

McGarrity, M. E., Sinn, D. L., & Gosling, S. D. (2015). Which personality dimensions do puppy tests measure? A systematic procedure for categorizing behavioral assays. *Behavioural Processes*, *110*, 117-124. https://doi.org/10.1016/j.beproc.2014.09.029

McMahon, E. K., Youatt, E., & Cavigelli, S. A. (2022). A physiological profile approach to animal temperament: How to understand the functional significance of individual differences in behaviour. *Proc Biol Sci*, *289*(1966), 20212379. https://doi.org/10.1098/rspb.2021.2379

Meagher, R. K. (2009). Observer ratings: Validity and value as a tool for animal welfare research. *Applied Animal Behaviour Science*, *119*(1), 1-14. https://doi.org/10.1016/j.applanim.2009.02.026

OpenAI. (2025). *ChatGPT* (9 Nov version) [Large language model]. https://chatgpt.com

R Core Team. (2025). *R: A Language and Environment for Statistical Computing*. In R Foundation for Statistical Computing. https://www.R-project.org/

Réale, D., Reader, S. M., Sol, D., McDougall, P. T., & Dingemanse, N. J. (2007). Integrating animal temperament within ecology and evolution. *Biological Reviews*, *82*(2), 291-318. https://doi.org/10.1111/j.1469-185X.2007.00010.x

Riemer, S., Müller, C., Virányi, Z., Huber, L., & Range, F. (2014). The predictive value of early behavioural assessments in pet dogs--a longitudinal study from neonates to adults. *PLoS ONE*, *9*(7), e101237. https://doi.org/10.1371/journal.pone.0101237

Rooney, N. J., Clark, C. C. A., & Casey, R. A. (2016). Minimizing fear and anxiety in working dogs: A review. *Journal of Veterinary Behavior*, *16*, 53-64. https://doi.org/10.1016/j.jveb.2016.11.001

Salmon, A., Driscoll, C., Paterson, M. B. A., Harpur, P., & Pachana, N. A. (2022). Issues Regarding the Welfare of Assistance Dogs. Animals, 12(23), 3250. https://doi.org/10.3390/ani12233250

Salonen, M., Mikkola, S., Niskanen, J. E., Hakanen, E., Sulkama, S., Puurunen, J., & Lohi, H. (2023). Breed, age, and social environment are associated with personality traits in dogs. *iScience*, *26*(5). https://doi.org/10.1016/j.isci.2023.106691

Serpell, J., Duffy, D. L., & Jagoe, J. A. (2016). Becoming a dog: Early experience and the development of behavior. In J. Serpell (Ed.), *The Domestic Dog: Its Evolution, Behavior and Interactions with People* (2 ed., pp. 93-117). Cambridge University Press. https://doi.org/10.1017/9781139161800.006

Serpell, J. A., & Hsu, Y. (2001). Development and validation of a novel method for evaluating behavior and temperament in guide dogs. *Applied Animal Behaviour Science*, *72*(4), 347-364. https://doi.org/10.1016/S0168-1591(00)00210-0

Sforzini, E., Michelazzi, M., Spada, E., Ricci, C., Carenzi, C., Milani, S., Luzi, F., & Verga, M. (2009). Evaluation of young and adult dogs' reactivity. *Journal of Veterinary Behavior*, *4*(1), 3-10. https://doi.org/10.1016/j.jveb.2008.09.035

Zeynivandnezhad, F., Rashed, F., & Kanooni, A. (2019). Exploratory Factor Analysis for TPACK among Mathematics Teachers: Why, What and How. *Anatolian Journal of Education*, *4*. https://doi.org/10.29333/aje.2019.416a

Appendix A

	Sub-test name	Description
1	Exploring	The dog is allowed 1 minute to explore the field area.
2	Walking	The handler walks to a fixed spot in the field and stands there for 15 seconds, without interacting with the dog.
3	Other side	The handler walks over to the other side of the field and stands there for 15 seconds, without interacting with the dog.
4	Calling	The dog is called by saying its name, at a distance of no more than 2 meters. The handler may clap to attract attention.
5	Course	The handler guides the dog through a obstacle course, and checks whether it wants to walk over certain surfaces and/or interact with certain objects.
6	Course with food	The course is repeated, and now the dog receives a food reward after each station.
7	Tidying up	The course is put away by the handler, producing a lot of noise.
8	Poncho	An unfamiliar person wearing a poncho forms a looming presence at a fixed spot in the field. The person removes the hood and faces the dog before they leave.
9	Self-control	Positioned next to the dog, the handler takes a few quick steps forward, then stops. Repeated three times.
10	Patience	The handler takes a step back after asking the dog to sit and giving a food reward. Repeated three times. Finally, the handler looks away and waits for the dog to seek interaction.
11	Noise	The handler distracts the dog while another person drops a metal box filled with marbles, a few metres behind the dog.
12	Chasing	The handler spins a flirt pole around the dog.

13	Umbrella	Opposite the dog and at eye-level, another person opens an umbrella.
14	Car	The handler moves a remote-controlled car back and forth across the field, repeated three times.
15	Veterinarian	The handler puts a leash on the dog, and touches its ears, mouth, paws, and tail.

Appendix B

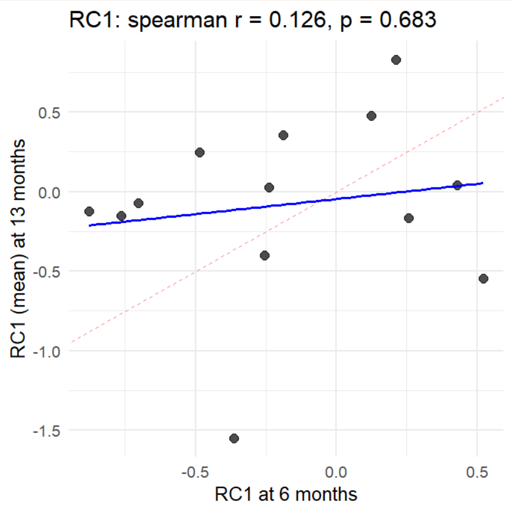
Category	**Behaviour**	**Type**	**Definition**	**Interpretation**
Orientation	*Orientation*	Duration	Head and/or body is turned towards the: Handler Stimulus Something else	Toward handler or something else: Stimulus avoidance Toward stimulus: Stimulus engagement
Tail	*Tail position*	Duration	Tail mostly low Tail partly low, partly medium to high Tail mostly medium to high	Low: Fearfulness Medium-High: Attentiveness
	Tail wagging	Duration	Movement of the tail from side to side, either: Relaxed/Excited, with broad strokes Tense, with narrow strokes	Relaxed/Excited: Attentiveness Tense: Fearfulness
	Tucked tail	Duration	No movement of tail as it is tucked tightly between hind legs, may be curled under ventral side.	Fearfulness
Vocalisations	*Barking*	Frequency	Clear, short and repetitive high-amplitude vocalisations directed toward stimulus. If continuous, score again after 3 s.	Excitability
	Growling	Frequency	Low, guttural vocalisations directed at: The stimulus The handler Nowhere in particular If continuous, score again after 3 s.	Aggression
	Whining	Frequency	Cyclic, low-amplitude vocalisation with mouth closed.	Fearfulness

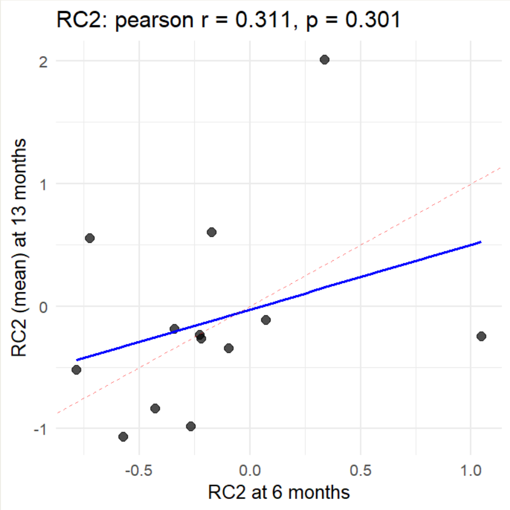
			If continuous, score again after 3 s.	
Locomotion	*Sitting*	Duration	Resting its body on the floor by extending its front legs and flexing its hind legs.	
	Standing	Duration	Standing upright by extending its legs against the floor.	
	Walking	Duration	Forward locomotion in a slow pace, slower than running.	
	Running	Duration	Forward locomotion in a rapid pace, faster than walking.	
Stimulus-related behaviours	*Approaching stimulus*	Frequency	Did not approach stimulus. Approached the stimulus within 1 body length after 5 s and within 30 s. Approached the stimulus within 1 body length within 5 s.	Stimulus engagement
	Proximity to stimulus	Duration	Time spent in contact with or within 1 body length from the stimulus. Time spent at 1 body length or further away from the stimulus.	Stimulus engagement
	Interacting with stimulus	Frequency	Pawing, grabbing or touching with its mouth, sniffing the stimulus.	Stimulus engagement
	Backing up	Frequency	Taking a few steps back, away from the stimulus.	Stimulus avoidance
	Fleeing	Frequency	Running away from the stimulus.	Stimulus avoidance
	Lunging	Frequency	Attempting to bite while moving in the direction of stimulus and/or jumping at it with fore paws, without physical contact.	Excitability

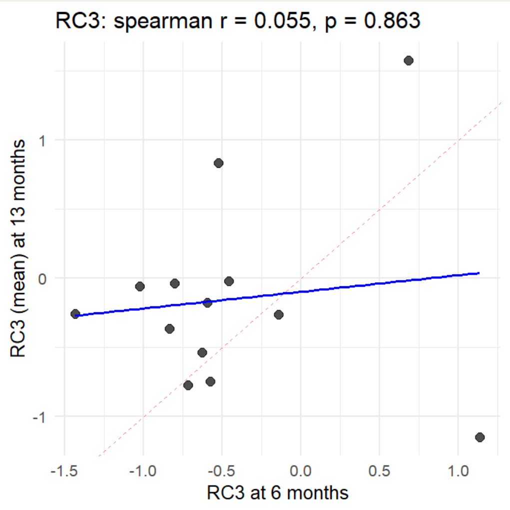
	Showing teeth	Frequency	Lips pulled back and revealing the teeth.	Aggression
Handler-related behaviours	*Approaching handler*	Duration	Did not approach the handler. Approached the handler within 1 body length.	Stimulus avoidance
	Proximity to handler	Duration	Time spent in contact with or within 1 body length from the handler. Time spent at 1 body length or further away from the handler.	Stimulus avoidance
Stress-related behaviours	*Trembling*	Duration	Visible vibrations of the entire body, excluding the tail.	Fearfulness
	Freeze	Duration	General rigidity of body posture, unmoving, with fixed attention toward the stimulus.	Stimulus avoidance
	Lip licking	Frequency	One or more brief movements of the tongue toward the nose or corners of the mouth.	Fearfulness
	Panting	Duration	Mouth open and tongue exposed with audible and/or observable heavy breathing.	Fearfulness
	Tucked ears	Duration	Ears folded against sides and/or back of head, having a flattened appearance.	Fearfulness
	Yawning	Frequency	Mouth open wide and inhaling.	Fearfulness

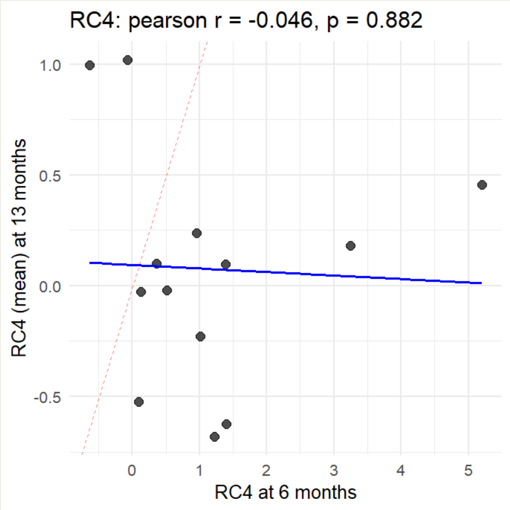
Appendix C1

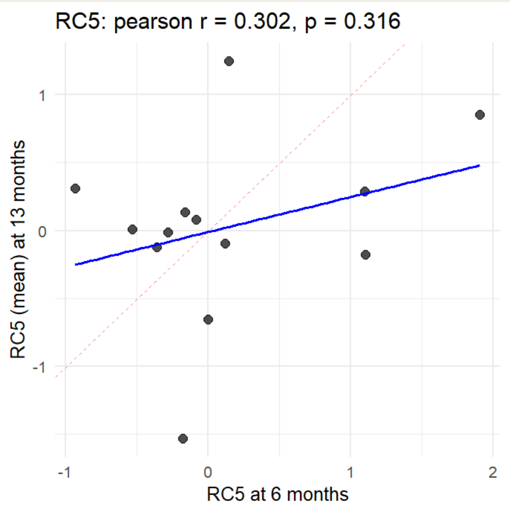
Plots of individual-level correlations between mean principal component scores at 6 and 13 months. The blue line is the trend line, the red-dotted line represents perfect correlation (x=y).





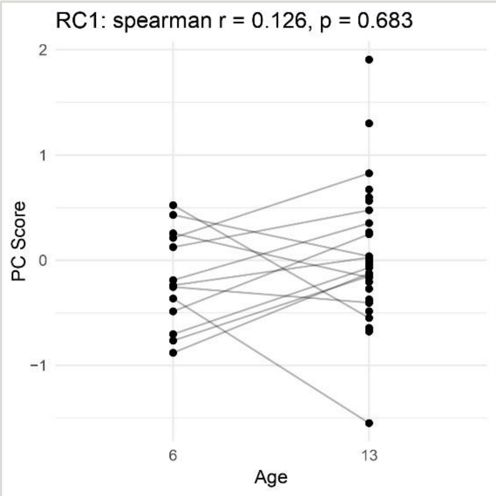


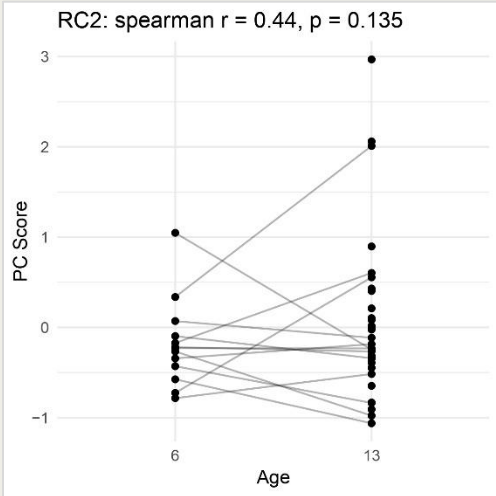


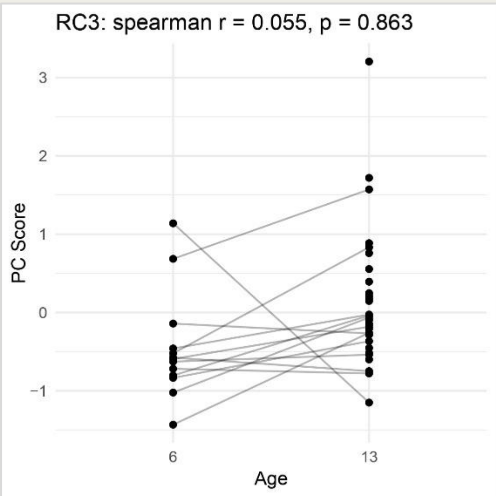


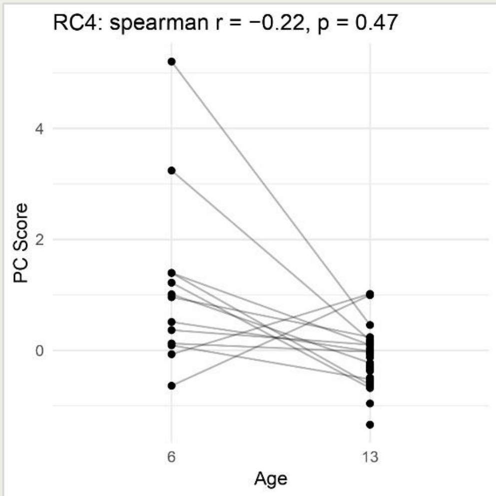
Appendix C2

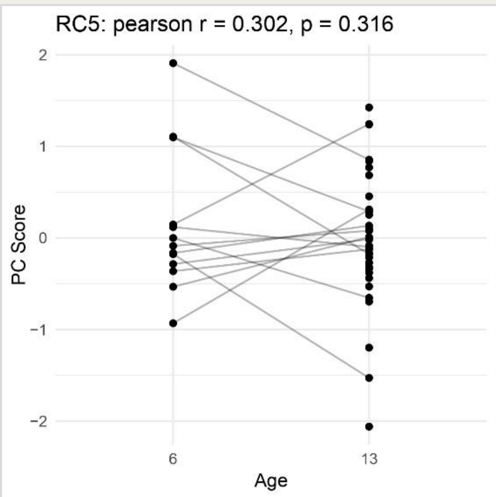
Plots of group-level correlations between mean principal component scores and age (6 and 13 months).





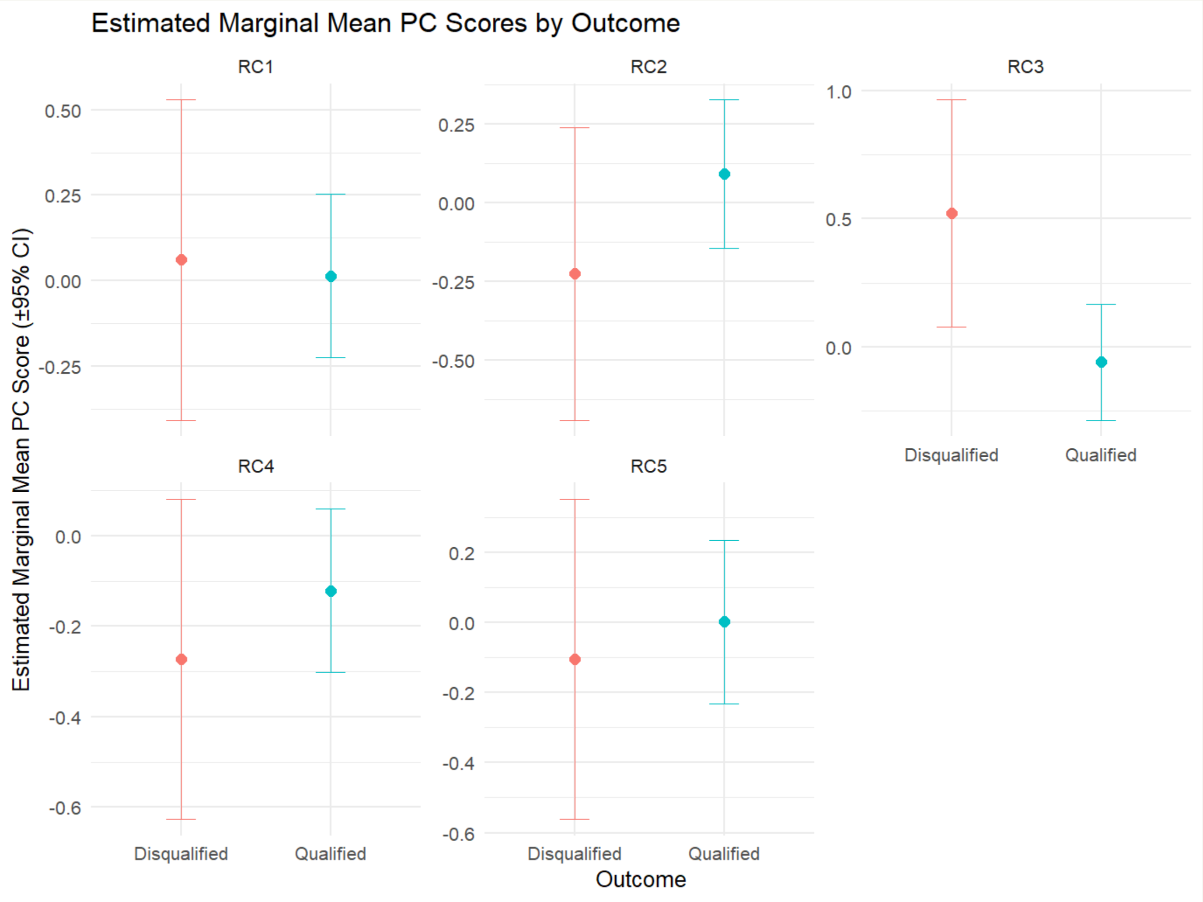






Appendix D

Plots of pairwise comparisons on principal component scores of qualified and disqualified assistance dog candidates.



Supplementary materials

- Appendix_E.xlsx
- R Script
- Covariance matrix